

# Linear Regression (Hồi quy tuyến tính)

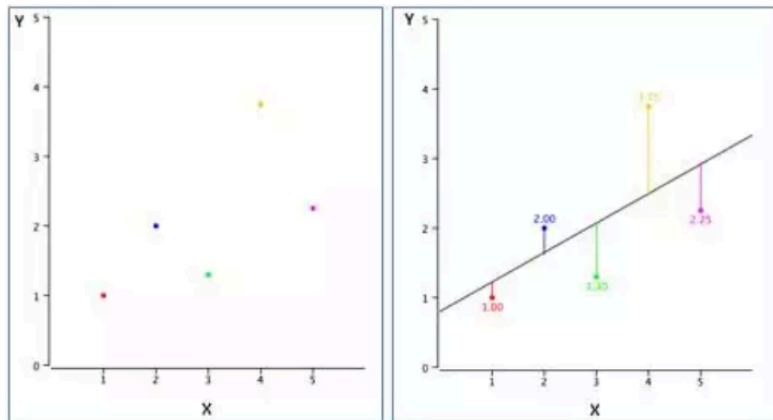
Hoàng Vũ Minh

Linear Regression hay hồi quy tuyến tính là một kỹ thuật học có giám sát, phân tích quan hệ giữa biến phụ thuộc Y với 1 hay nhiều biến độc lập x, dựa trên các dữ liệu có gán nhãn cho trước.

## 1. Ý tưởng cơ bản:

Hiểu một cách đơn giản bằng hình học: hồi quy tuyến tính tương đương với việc ta tìm một đường thẳng sao cho khoảng cách từ nó đến các điểm dữ liệu là nhỏ nhất (tức là nó sát với các điểm dữ liệu nhất), như vậy ta có thể dùng đường đó để dự đoán các điểm xuất hiện trong tương lai

Ví dụ trong trường hợp hồi quy tuyến tính đơn biến (chỉ có 1 biến):



Trong trường hợp đa biến, thay vì đường thẳng, nó sẽ là các mặt phẳng.

## 2. Phương trình cho trường hợp đơn biến:

$$Y_i = \beta_0 + \beta_1 * X_i$$

Trong đó:

- $Y_i$  là biến phụ thuộc
- $\beta_0$  là constant/intercept, hay hệ số tự do
- $\beta_1$  là slope (độ dốc), mô tả sự biến đổi của biến phụ thuộc Y.
- $X_i$  là các biến độc lập

Như vậy, ta cần tìm B0 và B1 sao cho giá trị dự đoán và giá trị thực tế sai khác ở mức tối thiểu. Phương pháp: OLS - bình phương nhỏ nhất có thể được sử dụng để tính toán ra B0 và B1 trong trường hợp này. Nghiệm khi sử dụng:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - B_1 \bar{x}$$

Với  $\bar{x}$  là các giá trị trung bình của x, được tính bằng:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

Với  $\bar{y}$  là các giá trị trung bình của y, được tính bằng:  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

### 3. Hàm mất mát:

Ta có thể sử dụng MSE (mean square error) làm hàm mất mát cho hồi quy tuyến tính. MSE sẽ đo trung bình bình phương sự chênh lệch giữa giá trị thực tế và giá trị dự đoán:

$$MSE = \frac{1}{n} \sum (Y_i - \hat{y}_i)^2$$

Trong đó  $\hat{y}$  là các giá trị dự đoán, y là các giá trị thực tế. Như vậy có thể nhận xét, MSE càng nhỏ thì giá trị dự đoán của mô hình càng sát với thực tế.

### 4. Độ đo cho các bài toán hồi quy

Cho các bài toán hồi quy, ta thường sử dụng metric RMSE và R2-square.

#### 4.1. RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_i - \hat{y}_i)^2}$$

RMSE có ý nghĩa tương tự với MSE, càng nhỏ thì chứng tỏ giá trị dự đoán càng sát với giá trị thực tế.

#### 4.2. R2

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Trong đó:

- $Y_i$  là giá trị thực tế
- $\hat{Y}$  là giá trị dự đoán
- $\bar{y}$  là giá trị trung bình của biến phụ thuộc

R2 đo lường tỉ lệ biến thiên của Y có thể được giải thích bằng X, nó nằm trong khoảng 0 đến 1. Tức là nếu R2 càng cao, càng gần 1 thì chứng tỏ biến phụ thuộc Y càng được biểu diễn tốt bằng x và ngược lại.

### **5. Nhược điểm của hồi quy tuyến tính:**

Mặc dù đơn giản và dễ sử dụng nhưng hồi quy tuyến tính tồn tại một số nhược điểm lớn:

- Nhạy cảm với các noise và outliers, những điểm này có thể làm cho kết quả của mô hình tệ đi rất nhiều, vì vậy để mô hình hoạt động tốt, cần xem xét loại bỏ các outliers
- Các biến độc lập có mối quan hệ tuyến tính mạnh với nhau: hiện tượng này có thể gọi là đa cộng tuyến, có thể gây ra ảnh hưởng rất lớn đến ước lượng hệ số hồi quy.
- Không diễn giải được các mối quan hệ phi tuyến

-> cách khắc phục các nhược điểm: tiền xử lý và kiểm tra trước khi xây dựng mô hình:

- Kiểm tra tính tuyến tính của dữ liệu, với trường hợp đơn biến, đơn giản là có thể vẽ đồ thị (plot) ra và xem nó có dạng đường thẳng hay không.

- Kiểm tra sự tương quan giữa các biến độc lập trước khi tiến hành hồi quy.